GISP@HEU'S SUBMISSION FOR TASK 1: HETEROGENEOUS AUDIO CLASSIFICATION IN THE DCASE 2026 CHALLENGE

Technical Report

Xiaoyu Feng[1], *Tong Ye*[1], *Xuefeng Yang*[1], *Feiyang Xiao*[1], *Qiaoxi Zhu*[2], *Jian Guan*[1*]

[1]Group of Intelligent Signal Processing, Harbin Engineering University, Harbin, China
[2]University of Technology Sydney, Ultimo, Australia

ABSTRACT

This technical report presents our submitted systems for Task 1: Heterogeneous Audio Classification in the Detection and Classification of Acoustic Scenes and Events (DCASE) 2026 Challenge. Our submission consists of four systems, including three individual systems and one ensemble system. System 1 and System 2 adopt a two-stage hierarchical training framework with hyperbolic representation learning, and are built upon different multimodal audio language backbones, i.e., Qwen2-Audio and Qwen3-Omni. System 3 is developed from the official baseline framework, using multimodal embeddings and hierarchical training with clean BSD10k data and filtered BSD35k data. System 4 is an ensemble system consisting of these three systems. Experimental results demonstrate that the proposed approach improves hierarchical classification performance and achieves 81.80% ± 0.21% on the development dataset.

***Index Terms*—** Heterogeneous audio classification, hierarchical learning, multimodal audio-language model, hyperbolic learning

1. INTRODUCTION

Heterogeneous audio classification aims to recognize diverse real-world sounds that vary widely in acoustic content, duration, recording conditions, and production style. In DCASE 2026 Challenge Task 1, this problem is formulated under the Broad Sound Taxonomy (BST) [1], a two-level taxonomy dataset consisting of 5 top-level categories and 23 second-level categories. The goal of the task is to predict the correct second-level BST label for each audio clip.

This task is challenging because the target sounds are highly heterogeneous and may include music, instrument samples, speech, sound effects, and soundscapes [2, 3, 4].

The conventional classifier may ignore the hierarchical relation between the top-level and second-level categories. For example, confusing two child classes under the same parent category is less severe than predicting a child class under a completely different parent category. Therefore, a model should not only optimize the second-level accuracy, but also preserve the top-level semantic structure of the taxonomy.

Recent audio-language models provide strong general-purpose acoustic and semantic representations and are promising backbones for heterogeneous audio classification [5, 6, 7, 8]. However, directly finetuning these models as flat 23-way classifiers may not fully exploit the hierarchical structure of BST, where parent-child relationships are important for both semantic modeling and evaluation.

Therefore, we propose a hierarchical finetuning framework based on pretrained audio-language models. Specifically, the finetuning process is decomposed into a parent-level adaptation stage and a joint parent-child optimization stage, enabling the model to first learn robust top-level representations and then refine second-level predictions under hierarchical guidance. We further employ hyperbolic learning to better model the hierarchical relationships between parent and child categories, together with ArcFace supervision [9] and contrastive learning [10, **?**] to improve category separation. Finally, the submitted prediction is produced by ensembling systems based on different multimodal audio-language backbones.

2. PROPOSED SOLUTION

2.1. Overall Framework

The goal of DCASE 2026 Task 1 is to predict the second-level label of each audio clip under the Broad Sound Taxonomy (BST) [1]. Since BST is organized as a two-level taxonomy, we formulate the task as a hierarchical classification problem instead of a flat 23-class classification problem. Given an input audio clip, the model is required to first capture its top-level semantic category and then determine the corresponding fine-grained second-level class.

Our framework is built upon pretrained multimodal audio-language models [6, 7, 8]. For each audio clip, an instruction-style prompt is constructed and fed into the audio-language backbone together with the audio signal. The model contains two prediction branches. The first branch performs parent-level classification over the 5 top-level BST categories, and the second branch performs second-level classification through parent-guided label generation. The optimization objective consists of a parent classification loss and a child generation loss. This design explicitly incorporates the hierarchical structure of BST into model optimization.

2.2. Two-stage Hierarchical Finetuning Strategy

To improve both top-level robustness and fine-grained discrimination, we adopt a two-stage hierarchical finetuning strategy. In the first stage, the pretrained audio-language model is adapted for parent-level classification using both BSD10k-v1.2 [2, 3] and BSD35k-CS [4], enabling the model to learn robust representations for the 5 top-level categories from a larger and more diverse training set. In the second stage, the model is further finetuned on BSD10k-v1.2 with both parent-level and second-level supervision, allowing it to refine fine-grained category discrimination while preserving

*Corresponding author.

the learned hierarchical structure. This training procedure allows the model to first establish reliable parent-level decision boundaries and then refine second-level predictions within the BST hierarchy.

2.3. Hyperbolic Representation Learning

The BST label space has an explicit hierarchical structure, where each second-level class belongs to one top-level parent category. Such a structure is naturally tree-like, which euclidean representations may be insufficient to model this hierarchical expansion efficiently [11, 12]. To better capture the hierarchical structure of BST, we introduce hyperbolic representation learning. Specifically, the parent-level representation extracted from the audio-language backbone is projected into a hyperbolic space before hierarchical discrimination. The hyperbolic representation is then used to enhance parent-level separation and preserve the semantic organization between parent and child categories.

To further improve representation quality, we incorporate ArcFace supervision and supervised contrastive learning. ArcFace enhances category discrimination by introducing an angular margin between classes, while supervised contrastive learning encourages samples from the same category to be closer and samples from different categories to be farther apart in the embedding space.Together, these objectives help the model learn more discriminative and hierarchy-aware representations for heterogeneous audio classification.

2.4. System Description

Our final submission consists of three systems that share the same hierarchical finetuning framework and hyperbolic representation learning strategy, while employing different pretrained audio-language backbones. All systems perform parent-level classification and parent-guided second-level prediction under the BST hierarchy.

System-1 is based on Qwen2-Audio, System-2 uses Qwen2.5-Omni, and System-3 adopts Qwen3-Omni. Despite the differences in backbone architecture, all systems follow the same training pipeline. System-4 is the ensemble system that integrates the outputs of System-1, System-2, and System-3. By combining predictions from models with different pretrained audio-language backbones, System-4 reduces backbone-specific errors and provides more robust final predictions.

3. EXPERIMENTS

3.1. Dataset

We conduct experiments on the DCASE 2026 Task 1 dataset. The task uses the Broad Sound Taxonomy, which contains 5 top-level classes and 23 second-level classes [2, 3, 4]. The final prediction target is the second-level class.

The dataset consists of two parts:

1. **BSD10k-v1.2**: a smaller manually annotated dataset with cleaner labels.
2. **BSD35k-CS**: a larger crowdsourced dataset with noisier labels.

BSD10k-v1.2 and BSD35k-CS are both derived from Freesound [13]. BSD10k-v1.2 is used as the main training and validation dataset for second-level classification. BSD35k-CS is used as auxiliary data in the parent-only training stage to improve coarse-level parent classification.

Table 1: Performance comparison with the official baseline on the development set.

System	Hier. Acc	Hier. F1
Baseline	79.71% ± 0.82%	78.76% ± 0.79%
System-1	81.13% ± 0.18%	80.83% ± 0.26%
System-2	82.91% ± 0.74%	81.34% ± 0.72%
System-3	81.07% ± 0.68%	80.97% ± 0.63%
System-4	82.44% ± 0.52%	81.80% ± 0.21%

3.2. Experimental Setup

For System-1, System-2 and System-3, the pretrained audio-language backbone is adapted to the BST classification task using Low-Rank Adaptation (LoRA) [14]. Training follows the two-stage hierarchical finetuning strategy described in Section 2. In the first stage, both BSD10k-v1.2 and BSD35k-CS are used for parent-level classification. In the second stage, the models are further finetuned on BSD10k-v1.2 with joint parent-level and second-level supervision.

During inference, the parent category is first predicted by the parent classification branch and then used to constrain the candidate set for second-level prediction. For the ensemble system, the outputs of the three individual systems are combined to produce the final prediction.

3.3. Evaluation Metrics

Following the official evaluation protocol [3], system performance is evaluated using hierarchical accuracy (Hier. Acc) and hierarchical F-score (Hier. F1). These metrics take the two-level structure of the BST into account by comparing the predicted and reference label paths in the hierarchy. The official ranking metric is the macro-averaged hierarchical F-score, which is computed by averaging the hierarchical F-scores across all second-level classes. This evaluation protocol rewards predictions that preserve hierarchical consistency for taxonomy-based audio classification tasks.

3.4. Results

We compare the proposed systems with the official baseline system of DCASE 2026 Task 1 on the development set. System-1 is based on Qwen2-Audio with hierarchical hyperbolic learning, while System-2 and System-3 use Qwen2.5-Omni and Qwen3-Omni as the backbone, respectively. System-4 is the ensemble of the three individual systems. All of our systems outperform the official baseline in terms of hierarchical F-score, as shown in Table 1.

4. CONCLUSION

In this technical report, we introduce our submission systems for DCASE 2026 Challenge Task 1. Our method combines a two-stage hierarchical finetuning strategy with hyperbolic representation learning, and further ensembles multiple multimodal audio-language backbones to produce the final prediction. The experiments show that the proposed systems outperform the official base-

line and improve hierarchical classification performance for heterogeneous audio classification.

5. ACKNOWLEDGMENT

This work was partly supported by the Heilongjiang Provincial Natural Science Foundation of China under Grant No. BS2025F009.

6. REFERENCES

[1] P. Anastasopoulou, X. Serra, and F. Font, "A general-purpose sound taxonomy for the classification of heterogeneous sound collections," In press.

[2] P. Anastasopoulou, J. Torrey, X. Serra, and F. Font, "Heterogeneous sound classification with the Broad Sound Taxonomy and Dataset," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2024.

[3] P. Anastasopoulou, F. A. Dal Rí, X. Serra, and F. Font, "Hierarchical and multimodal learning for heterogeneous sound classification," in *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*, 2025.

[4] P. Anastasopoulou and F. Font Corbera, "Bsd35k-cs (broad sound dataset 35k - crowd sourced)," March 2026.

[5] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "CLAP: Learning audio concepts from natural language supervision," 2022.

[6] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin, C. Zhou, and J. Zhou, "Qwen2-Audio technical report," 2024.

[7] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin, "Qwen2.5-Omni technical report," 2025.

[8] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu, Y. Lv, Y. Wang, D. Guo, H. Wang, L. Ma, P. Zhang, X. Zhang, H. Hao, Z. Guo, B. Yang, B. Zhang, Z. Ma, X. Wei, S. Bai, K. Chen, X. Liu, P. Wang, M. Yang, D. Liu, X. Ren, B. Zheng, R. Men, F. Zhou, B. Yu, J. Yang, L. Yu, J. Zhou, and J. Lin, "Qwen3-Omni technical report," 2025.

[9] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proc. the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[10] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *Proc. Advances in Neural Information Processing Systems(NeurIPS)*, vol. 33, pp. 18 661–18 673, 2020.

[11] M. Nickel and D. Kiela, "Poincaré embeddings for learning hierarchical representations," in *Proc. Advances in Neural Information Processing Systems(NeurIPS)*, vol. 30, 2017.

[12] O.-E. Ganea, G. Bécigneul, and T. Hofmann, "Hyperbolic neural networks," in *Proc. Advances in Neural Information Processing Systems(NeurIPS)*, vol. 31, 2018.

[13] F. Font, G. Roma, and X. Serra, "Freesound technical demo," in *Proc. ACM International Conference on Multimedia*, 2013, pp. 411–412.

[14] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. International Conference on Learning Representations*, 2022.